**Trabajo Final PySpark**

**Aplicación de Algoritmos de Clasificación en PySpark sobre Dataset Real**

**Instrucciones generales**

* Utilizar PySpark para todo el procesamiento de datos.
* Completar cada sección solicitada.
* Utilizar alguno de los datasets recomendados (ver abajo).
* Puntaje total: **39 puntos**.

**Parte 1: Carga y Exploración de Datos (3 puntos)**

* Cargar el dataset en una sesión de PySpark.
* Mostrar primeras filas, esquema, y estadísticas descriptivas.
* Contar valores faltantes por columna.

**Parte 2: Preprocesamiento de Datos (4 puntos)**

* Imputar valores nulos:
  + Variables numéricas: media o mediana.
  + Variables categóricas: "Sin Informacion".
* Codificar variables categóricas.

**Parte 3: Separación de Datos (2 puntos)**

Dividir en conjunto de entrenamiento (80%) y prueba (20%).

**Parte 4: Entrenamiento de Modelos (20 puntos)**

* Entrenar **mínimo 2 algoritmos** entre:
  + Random Forest
  + Decision Tree
  + Naive Bayes
* Mostrar las métricas:
  + Accuracy
  + Precision
  + Recall
  + F1 Score

**Parte 5: Visualización de Matriz de Confusión (5 puntos)**

* Graficar la matriz de confusión para cada modelo.
* Explicar con sus palabras el análisis de la matriz de confusión.

**Parte 6: Análisis Final y Conclusiones (5 puntos)**

* ¿Qué modelo tuvo mejor desempeño?
* ¿Qué dificultades encontraron en el flujo de trabajo?
* Las conclusiones deben ser entregadas en un archivo .pdf

**Sugerencia de Datasets Kaggle (usar uno)**

* [Heart Failure Prediction Dataset](https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction)
* [Stroke Prediction Dataset](https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset)
* [Diabetes Health Indicators Dataset](https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset)

**Notas Finales**

* El trabajo debe estar ordenado, bien comentado y ejecutable de principio a fin.
* Entregar el dataset trabajado
* No eliminar columnas sin justificación clara.
* No se aceptan trabajos fuera del plazo 2/05/2025.